Intercontinental Geoinformation Days

igd.mersin.edu.tr

Determination of effective predisposing factors using random forest-based gini index in landslide susceptibility mapping

Alihan Teke*1, Taskin Kavzoglu 1

1Gebze Technical University, Department of Geomatics Engineering, Kocaeli, Turkey

Keywords
Landslide susceptibility
Machine learning
Random forest
Feature importance
Factor selection

ABSTRACT

Landslide susceptibility mapping is a multi-phase procedure that includes several key steps, one of which is the correct determination of landslide predisposing factors. In the current literature, however, there is no global consensus or framework about the selection of these factors. In this study, the effectiveness of predisposing factors was investigated using the Random Forest-based Gini index to generate landslide susceptibility models. For this purpose, 16 predisposing factors, representing the morphological, lithological, and environmental characteristics of the study area, were initially utilized and measured their importance scores calculated by utilizing the Gini index. Then, three models (RF-1, RF-2, and RF-3) including 50%, 75%, and the whole of the factors, were produced based on the importance scores. To select the optimum one among these models, their performances were assessed employing two accuracy assessment metrics, namely overall accuracy (OA) and area under curve (AUC). The validation results revealed that AUC obtained using RF-1, RF-2, and RF-3 models were calculated as 85.85%, 96.70%, and 90.66% respectively. Also, the statistical significances of the models were evaluated using McNemar's test, which revealed that all models were statistically different from each other.

1. INTRODUCTION

At a global scale but particularly in mountainous zones, landslides have drastically shaped and modified the local terrain due to the deformations they create (Geertsema et al. 2009). As a result of these surface displacements, both human life and man-made structures including settlements, infrastructure, superstructure, and also ecological integrity are under threat. In addition to these adverse influences, landslides also lead to tremendous economic damages and the devastation of natural resources (Schuster & Fleming, 1986). Therefore, implementing the necessary preparedness strategies and generating emergency action plans are of utmost significance in the pre-disaster phase given all the above-mentioned issues (Gómez & Kavzoglu, 2005). In this context, landslide susceptibility maps are highly functional tools that can display the distribution of landslide-susceptible and non-landslide zones and enable the reactivation of idle fields. Thus, landslide susceptibility maps provide valuable supports to public enterprises in terms of both preventing financial losses and safeguarding human beings (Kavzoglu et al. 2014).

Given the complex dynamics of landslides, the performances of landslide susceptibility maps depend primarily on the determination of convenient and robust landslide predisposing factors. In the literature, generally, the selection of these factors is executed with the aid of expert judgment or utilizing available geo-environmental data sets (e.g. lithology map, road networks). However, some factors in the data set may reduce the accuracy of models, which lead to adversely affect the reliability of the produced landslide susceptibility maps. Furthermore, unnecessary features not only increase the computational complexity of the model, but also extend the processing time.

To alleviate the aforementioned challenges, the feature selection paradigm has been recently gained popularity as an effective solution. Broadly speaking, FS approaches discard superfluous variables from the data set, and thus, more comprehensible models are generated. Consequently, both the predictive performance and reliability of the landslide susceptibility maps can be enhanced. Due to these wide-ranging benefits, feature selection algorithms have been

* **Corresponding Author**

* (a.teke2020@gtu.edu.tr) ORCID ID 0000 – 0003 – 4048 – 329X
(kavzoglu@gtu.edu.tr) ORCID ID 0000 – 0002 – 9779 – 3443

Cite this study

Teke A & Kavzoglu T (2021). Determination of Effective Predisposing Factors Using Random Forest-Based Gini Index in Landslide Susceptibility Mapping. 2nd Intercontinental Geoinformation Days (IGD), 198-201, Mersin, Turkey

intensively employed in landslide susceptibility zonation studies (Hu et al. 2021; Kavzoglu et al. 2015; Sahin 2020).

The main motivation of this current work is to determine the optimal landslide predisposing factor subset using feature importance scores obtained by the Gini index of Random Forest (RF). To achieve this object, three models consisting of 50%, 75%, and the whole of the factors were established based on the importance order of 16 landslide causative factors. To determine the optimal subset of factors, two accuracy assessment metrics, namely overall accuracy (OA) and area under curve (AUC) were calculated. Finally, McNemar's test was employed to statistically analyze the performance differences of the three models.

2. METHOD

2.1. Study Area and Datasets

The Arakli district of Trabzon province situated in the northeast part of Turkey, between latitudes 40° 57′ N and 40° 31′ N; longitudes 39° 49′ E and 40° 04′ E, was selected as the region of interest (Fig. 1). It covers an area of about 479 km^2, the large part of the study area is mountainous with elevations varying 0 to 2876m and slopes up to about 70°. In addition to the morphological characteristics, heavy rainfall is one of the most significant agents that predispose to the occurrence of landslides in the study area (Sahin et al. 2017). Furthermore, anthropogenic influences such as the construction of superstructure, infrastructure and deforestation make a substantial contribution to the occurrence of mass movement.

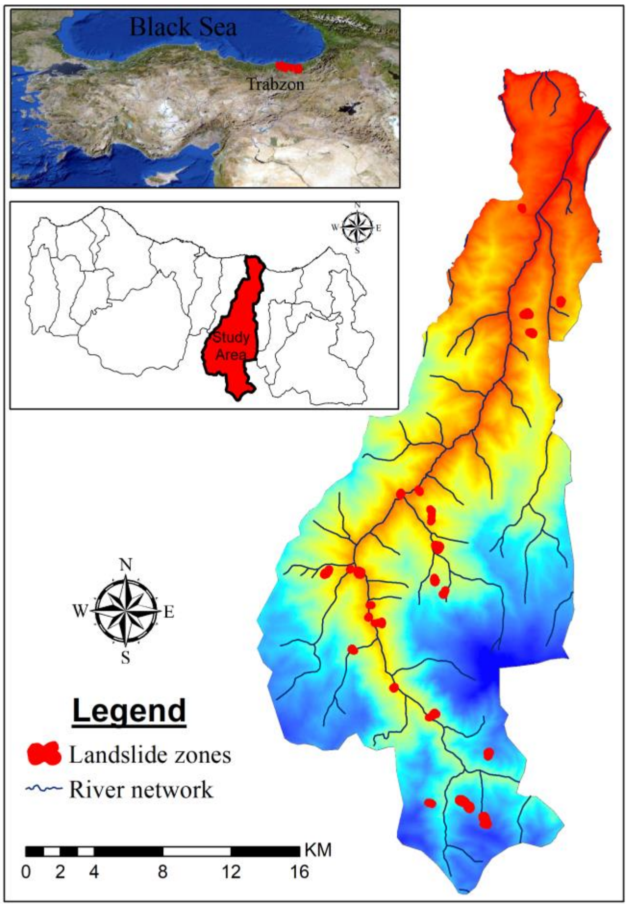

Figure 1. Location map of the study area and landslides

In the process of reliable and accurate generation of landslide susceptibility maps, one of the key steps is the preparation of the landslide predisposing factors (Kavzoglu et al. 2020). In this study, slope, topographic roughness index (TRI), elevation, distance to roads, distance to rivers, profile curvature, aspect, valley depth, normalized difference vegetation index (NDVI), lithology, topographic wetness index (TWI), stream power index (SPI), road density, topographic position index (TPI), drainage density, and plan curvature were determined as primary landslide predisposing factors to generate landslide susceptibility map.

2.2. Random Forest

Random forest (RF), proposed by Breiman (2001), is a solid ensemble learning approach whose working principle is essentially based on combining many decision trees. In addition to its capability to perform versatile tasks such as classification, regression, and unsupervised learning, RF has been extensively employed for feature selection applications. It reduces the variance and enhances the generalization, resulting from the bagging methodology. To assess the predictive performance of the model, about two of thirds of the instances or simply in-bag instances are employed for the model training phase while remaining (i.e. out-of-bag) instances are employed for the internal cross-validation procedure (Kavzoglu et al. 2018).

In the current literature, several metrics, which are relatively cost-effective, have been proposed to calculate factor importance scores using RF (Fabris et al. 2018). Among them, the Gini index (or mean decrease impurity) is widely preferred by researchers to measure the importance of features. Gini index is essentially a splitting function used by the RF algorithm in order to specify which feature to split on throughout the learning stage (Qi 2012).

In the domain of the earth sciences and geomorphological researches, the RF technique has become quite popular and attracted wide interest owing to its unique abilities in coping with complicated and inconsistent real-world problems. From the perspective of landslide susceptibility mapping studies, recent works have demonstrated a broad application of the RF (Merghadi et al. 2020). Although RF algorithms have been utilized for regression and classification purposes, studies in which RF is employed as a feature selection algorithm are rare to find in the field of landslide susceptibility mapping.

3. RESULTS

In this study, a RF-based feature selection algorithm was adopted to detect the most efficient landslide contributing factors. In parallel with this purpose, 16 factors were initially included in the modeling process. Then, the importance value of each factor was calculated using the Gini index, as shown in Table 1. Results revealed that the slope had the greatest importance score of 0.358. Other significant factors were determined as TRI (0.222) and elevation (0.100). On the other hand, plan curvature (0.009), drainage density (0.015), and TPI (0.018) were found to be the least effective ones.

Considering the importance scores, three models (RF-1, RF-2, and RF-3) including 75%, 50%, and the whole of the factors were produced (Table 1). Whereas the RF-1 model had all factors in the data set, RF-2 and RF-3 factors had %75 and %50 of the data set, respectively. Among these models, to seek the best ones, each model was independently trained using the RF classifier.

Table 1. Three models formed with different factors based on factors importance scores

Factors	RF-1	RF-2	RF-3	Importance Score
Slope	√	√	√	0.358
TRI	√	√	√	0.222
Elevation	√	√	√	0.100
Distance to Roads	√	√	√	0.038
Distance to Rivers	√	√	√	0.035
Profile Curvature	√	√	√	0.032
Aspect	√	√	√	0.030
Valley Depth	√	√	√	0.028
NDVI	√	√	X	0.027
Lithology	√	√	X	0.027
TWI	√	√	X	0.027
SPI	√	√	X	0.027
Road Density	√	X	X	0.025
TPI	√	X	X	0.018
Drainage Density	√	X	X	0.015
Plan Curvature	√	X	X	0.009

Evaluation of performances has been considered as an essential tool in obtaining information about the reliability of landslide susceptibility maps. Therefore, two accuracy assessment measures including OA and AUC were calculated in order to compare the predictive performances of three models. According to the results of the accuracy assessment, the RF-2 model had the greatest OA with 93.28% and followed by the RF-3 and RF-1 models with 83.48% and 75.02, respectively (Table 2). Similarly, when it comes to the AUC value, the RF-2 model had the highest AUC value 96.70% followed by the RF-3 (90.66%) and RF-1 (85.85) models, as shown in Fig. 2.

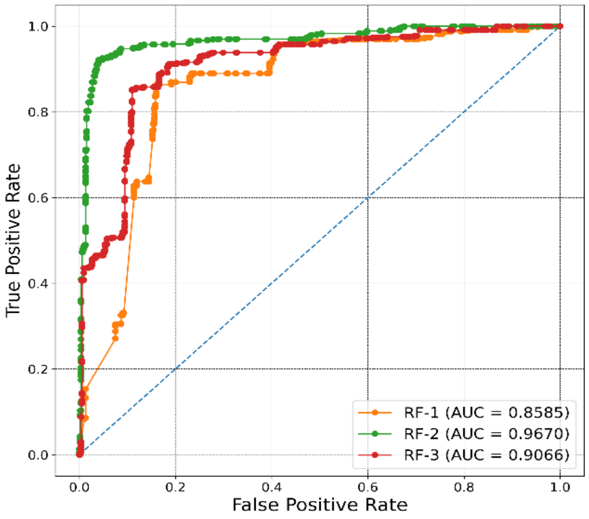

Figure 2. ROC curve analysis of the models

Table 2. Performance evaluation of the models

Models	OA (%)	AUC (%)
RF-1	75.02	85.85
RF-2	93.28	96.70
RF-3	83.48	90.66

Table 3. McNemar's test results for the models

	RF-1	RF-2	RF-3
RF-1	-	154.58	72.75
RF-2		-	78.22
RF-3			-

Apart from these accuracy assessment measures, the statistical significance of the difference between the performances of the models was also analyzed using McNemar's test. If the estimated statistical value is higher than the chi-square table value (3.84 for a 95% confidence interval), it can be inferred that the difference between the results of the two independent models is statistically significant. In other words, the model outperforms the other model. According to the estimated statistical significance test results, it was observed that three models yielded statistically different results. More specifically, when analyzed the statistical significance of the RF-1 and RF-2 models, the statistical value is estimated as 154.58, which clearly indicates statistical significance in performances.

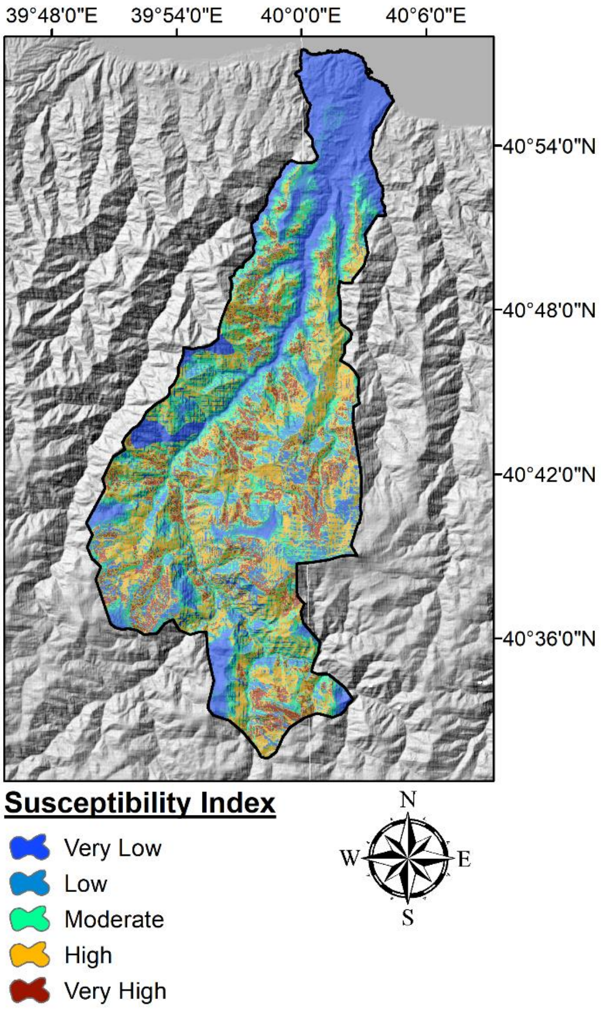

Figure 3. Landslide susceptibility map produced using optimal landslide predisposing subset including 12 factors

4. CONCLUSION

In this current study, the effectiveness of predisposing factors was analyzed using the RF-based Gini importance scores to create a more robust and accurate landslide susceptibility map for the Arakli district of Trabzon Province, Turkey. For this purpose, three models consisting of 50%, 75%, and the whole factors in the data set were produced based on the importance order of features obtained by utilizing the Gini index algorithm. Two accuracy assessment (OA and AUC) and a statistical significance test was used to make a sound comparison of the model performances.

According to the indication of the study, some considerable inferences can be drawn. Firstly, it was observed that the RF-2 model consisting of 12 landslide causative factors (slope, TRI, elevation, distance to roads, distance to rivers, profile curvature, aspect, valley depth, NDVI, lithology, TWI, and SPI) were found to be more efficient than other models. Thus, 25% of the whole data set was curtailed, which also alleviated the training time and model complexity. Besides, it can be clearly concluded that using whole conditioning factors caused the curse of dimensionality, also called the Hughes phenomenon. When utilized RF-3 model containing 50% of the whole data set, a significant decrease in the model performance by about 10% was observed. As a result, the RF-2 model with 93.28% OA, which is determined as the optimal model, was employed in the generation of the landslide susceptibility map of the study area. Secondly, the slope was found to be the most important causative factor, which is compatible with many previous types of research. Thirdly, it can be observed that RF algorithms employed with the correct features were yielded highly accurate results. This indication proves the robustness of the RF algorithm. In summary, the correct identification of predisposing factors, which is one of the critical issues in the determination of landslide susceptibility, was carried out utilizing the RF-based Gini index algorithm and the landslide susceptibility map with high accuracy and reliability was attained.

REFERENCES

Breiman L (2001). Random forests. Machine Learning, 45(1), 5-32.

Fabris F, Doherty A, Palmer D, De Magalhães JP & Freitas AA (2018). A new approach for interpreting Random Forest models and its application to the biology of ageing. Bioinformatics, 34(14), 2449-2456.

Geertsema M, Highland L & Vaugeois L (2009). Environmental impact of landslides. Landslides-Disaster Risk Reduction, 1, 589-607.

Gómez H & Kavzoglu T (2005). Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin, Venezuela. Engineering Geology, 78(1-2), 11-27.

Hu X, Mei H, Zhang H, Li Y & Li M (2021). Performance evaluation of ensemble learning techniques for landslide susceptibility mapping at the Jinping county, Southwest China. Natural Hazards, 105(2), 1663-1689.

Kavzoglu T, Colkesen I & Sahin EK (2018). Machine Learning Techniques in Landslide Susceptibility Mapping: A Survey and a Case Study. Landslides: Theory, Practice and Modelling. Advances in Natural and Technological Hazards Research, 50,283-301.

Kavzoglu T, Sahin EK & Colkesen I (2015). Selecting optimal conditioning factors in shallow translational landslide susceptibility mapping using genetic algorithm. Engineering Geology, 192, 101-112.

Kavzoglu T, Sahin EK & Colkesen I (2014). Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. Landslides, 11, 425-439

Kavzoglu T, Teke A & Bilucan F (2020). Effectiveness of machine learning algorithms in landslide susceptibility mapping: A case study of Trabzon Province, Turkey. 41st Asian Conference on Remote Sensing.

Merghadi A, Yunus AP, Dou J, Whiteley J, ThaiPham B, Bui DT, Avtar R & Abderrahmane B (2020). Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. Earth-Science Reviews, 207, 103225.

Qi Y (2012). Random forest for bioinformatics. Ensemble Machine Learning: Methods and Applications, Springer (2012), 307-323.

Sahin EK (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. SN Applied Sciences, 2, 1-17.

Sahin EK, Ipbuker C & Kavzoglu T (2017). Investigation of automatic feature weighting methods (Fisher, Chi-square and Relief-F) for landslide susceptibility mapping. Geocarto International, 32(9), 956-977.

Schuster RL & Fleming RW (1986). Economic losses and fatalities due to landslides. Bulletin of the Association Engineering Geologists, 23(1), 11-28.